古文 LIWC 词典的构建及初步分析

范妙榕[1,2] 邢付贵[1,2] 刘兴云[1,2] 朱廷劭[2*]
[1]（中国科学院大学 北京 100049）
[2]（中国科学院心理研究所，北京 100101）

摘要：[背景]LIWC（基于语词计量的文本分析）以关键词的词频统计为基础，可对个体和群体的表达语句的心理学意义等方面进行量化分析。由于文言文的表达方式与现代汉语存在明显的差异，为了分析文言文文本的心理学意义，我们在简体中文 LIWC 词典（Simplified Chinese LIWC 2015 年版本，简称 SC-LIWC)的基础上，构建了古文 LIWC(Classical Chinese LIWC，以下简称 CC-LIWC）词典。[目的]本研究的目的是探究如何构建 CC-LIWC 词典并介绍如何使用该词典对古文文本进行分析。[方法]获取在线汉语词典的全部词汇及其对应解释，保留文言文词及其现代文译文，并从译文中寻找 SC-LIWC 词，将 SC-LIWC 词与文言文词进行匹配。对匹配结果进行人工标注，确保结果的一致性与准确性。[结果]最终生成的 CC-LIWC 包含了 81 个词类与 49136 个文言文词条。[局限]古文中一词多义、一词多性的情况较为普遍，对词典中词汇的分类存在一定影响。[结论]使用 CC-LIWC 对《论语（节选）》、《孤愤》进行词频分析，分析结果体现了儒家的中庸与法家的注重逻辑辩证的区别，说明 CC-LIWC 词典能够有效区分文本的表达倾向。

关键词： 词典；文言文；LIWC；词频统计
分类号：B849

Classical Chinese LIWC: A Brief Introduction and Pilot Analysis

FanMiaorong[1,2] XingFugui[1,2] LiuXingyun[1,2] ZhuTingshao[2*]
[1]（University of Chinese Academy of Sciences, Beijing 100049）
[2] (Institute of Psychology，Chinese Academy of Sciences Beijing 100101）

Abstract:
[Background] Based on counting frequency of specially selected words, LIWC (known as Linguistic Inquiry and Word Count) is a useful tool to analyze expressions of writings or other texts created by individuals or group, for purpose of figuring out the psychological meanings inside the texts. In ancient China, the classical style of writing has a striking difference with modern times. In order to analyze the psychological meanings of classical Chinese text, we construct a Classical Chinese version of LIWC dictionary (known as CC-LIWC), based on the 2015 edition of Simplified Chinese LIWC (known as SC-LIWC).
[Objective] In this paper, we show the constructing process of CC-LIWC and give an example of how to use the dictionary to analyze classical Chinese text.
[Methods] First, we obtain all the words (including modern Chinese and Classical Chinese words) and their corresponding explanations from the online Chinese dictionary and keep the classical Chinese words with their modern translation; second, we search SC-LIWC words in the explanations. In this way, SC-LIWC words are mapping with the classical Chinese words; finally,

we invite ancient Chinese based professionals to check the mapping results manually to ensure the consistency and accuracy of the results.
[Results] The final dictionary includes 81 categories and 49136 classical Chinese entries.
[Limitations] In classical Chinese context, polysemy or diversity of a word is very common, which affects the classification of words in the dictionary.
[Conclusion] we use CC-LIWC to analyze *The Analects(excerpts)* and *The Isolated Indignation*. The result shows the difference between the moderation of Confucian and the dialectical thinking of Legalist. Therefore, CC-LIWC dictionary can distinguish the expression tendency of text efficiently.
Key words: Dictionary；Classical Chinese；LIWC；Word Frequency

1 引言

现有对中国历史的研究，多从定性层面上进行分析与解读，而鲜有研究从量化分析角度开展实证研究。随着大数据及自然语言处理技术（Natural Language Process, NLP）的发展与日益成熟，我们现在可以借助数理统计与计算机技术，对中国历史史料记载等进行量化分析，从而佐证以往的定性分析结论或得出更多不同以往的新论点。此外，利用大数据技术，我们还可以处理更大规模的历史文本数据，突破以往人工研究的局限性，从一个更为宏观且实证的角度去看待中国历史演化过程。历史是由人创造的，对历史的分析，离不开对生活在那个历史年代的人们的分析，包括群体与个体，文化与心理等。利用现代信息技术，可以基于史料数据的基础，进行很多方面的全新研究，开展数字化心理考古，进行共时性分析（横向对比各朝代的不同特点）、历时性分析（特定文化的时代变迁）、对特定群体或个体的心理特征进行分析（从科学心理学层面建立对中国历史人物的人格解读）等。

LIWC 词典是一种基于语词计量的文本分析工具，其开发的目的在于使用计算机程序代替人工评分对文本进行分析，其用途主要是对个体和群体的表达语句的心理学意义等方面进行量化分析。LIWC 是属于自然语言处理技术中的一种，它可以对文本内容进行量化分析并将导入的文本文件的不同类别的词语（尤其是心理学类词语）加以计算，比如因果词、情绪词、认知词等心理词类在整个文本中的使用百分比[1]。LIWC 主要是统计文本中反映不同情绪、认知过程、个体等类别的词所占文本的百分比，从量化角度理解文本所表达的内容，而不需人工的参与，适用于对大量文本进行量化分析的场景。

一个完整的 LIWC 包含 LIWC 词典主体及对应的词频统计程序。其中 LIWC 词典包含两个部分，第一个部分是 LIWC 类别编号及类别名称，第二部分是 LIWC 词汇及其所属类别编号。

不同于现代汉语，文言文作为古人写作用的文体，有其特定的表达方式。为了分析文言文文本的心理学意义，我们在 SC-LIWC 词典的基础上，构建了古文 LIWC 词典(Classical

Chinese LIWC, CC-LIWC)词典。本研究的目的是探究如何构建 CC-LIWC 词典并介绍如何使用 CCLIWC2015 对文言文文本进行初步分析。

2 方法

本文采用的翻译方法是先找到文言文词及其现代译文，从译文中查找对应的 SC-LIWC 词，将 SC-LIWC 词与文言文进行匹配。通过这种方式将所有 SC-LIWC 现代汉语词反向翻译为文言文词汇。我们邀请了古文专业的研究生通过人工标注的方式确认文言文候选词与 LIWC 词的意思一致，确保翻译的准确性。

2.1 数据收集

本研究采用的数据包括文言文词及 LIWC 词典两个部分：（1）文言文词部分，来源于在线汉语词典[3]的词汇及其释义与示例；（2）LIWC 词典部分，采用了 SCLIWC2015 版本。

2.2 数据处理

对于从在线汉语词典获取到的词汇，单字词使用文言文字典[4]进行了过滤，双字词及多字词，使用汉语古典文本数据库[5]进行过滤，保留古汉语词及其译文，删除非古汉语词汇，保留共计 80 多万条词条。

对于 SCLIWC2015 词典，由于存在部分英文单词与网络用语等，在古代没有对应语境，因此我们将现代汉语特有（文言文中无对应语义）的词汇剔除。类别部分保留了原有的 81 个词类，词表部分去掉了一些网络词和标点符号等，一共去掉了 1365 个词，留下 8438 个词用于翻译。

最后从古汉语词的译文中寻找 SC-LIWC 词，由此将 SC-LIWC 词与古汉语词进行匹配，形成对应关系。

2.2.2 人工标注

我们邀请了 6 位古文专业研究生对匹配结果进行人工标注。目的在于确认 SC-LIWC 词汇与其匹配的古汉语词意思一致，确保匹配的准确性。待标注数据分为两批，每批分为两组数据，共计 4 组数据。为了保证标注结果的一致性，我们对标注要求进行了基本的解释说明，取出小部分数据给标注者进行标注，测试标注者之间标注的一致性，确认一致性达到 85%以上，才开始进行正式标注。

2.2.3 二次核查

标注完成后，对于所有标注数据进行汇总，对一次标注的数据进行二次核查。目的在于检查标注结果是否正确以及去除重复翻译的词汇。

3 结果

通过词典搜索以及后续的人工标注，我们最后生成 CC-LIWC 词典包含 81 个词类与 49136 条词条。

（1）词汇词类分布情况

每个大词类的词汇数量分布情况如图 1 所示（按照 CC-LIWC 中词汇数占比从高到低排序）。CC-LIWC 词典中，词类词数占比排名前五的大类为：情感词、动机词、个人关切、相对词、认知过程词。相比 SC-LIWC，CC-LIWC 词典中收录的情感词、认知过程词、动机词比例明显增加，而功能词、生理过程词、个人关切词和非正式语言比例明显减少。

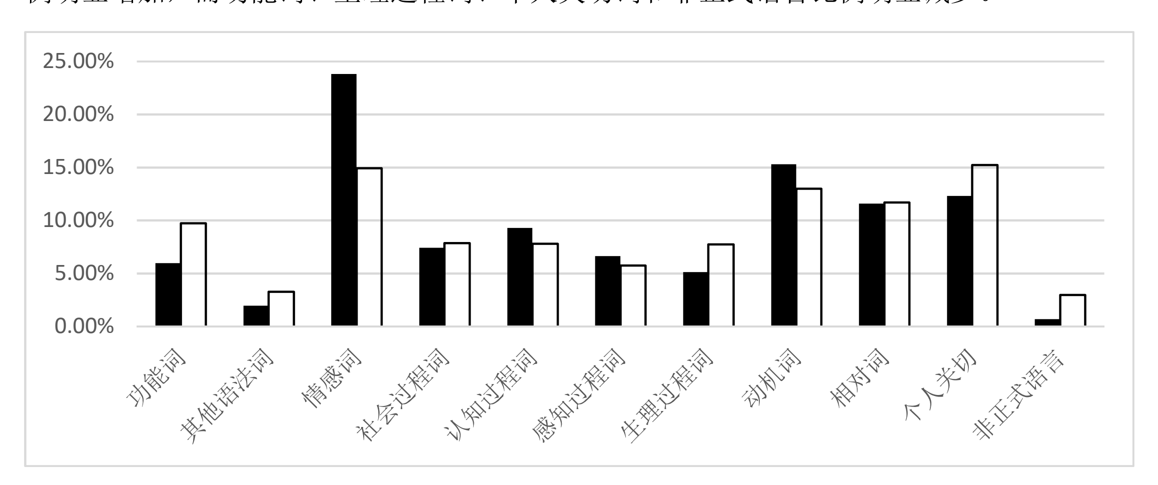


图 1 CC-LIWC（左）与 SC-LIWC（右）词类词汇数分布情况对比

（2）单字词与多字词分布情况

如表 1 所示，CC-LIWC 中单字词与双字词占比降低，而多字词占比增高。

表 1 CC-LIWC 与 SC-LIWC 单字词、多字词词数占比情况对比

	CC-LIWC		SC-LIWC	
	词数	占总词数比例	词数	占总词数比例
单字词	2566	5.22%	946	9.73%
双字词	37287	75.89%	7567	77.85%
多字词	9283	18.89%	1207	12.42%
总词数	49136	100.00%	9720	100.00%

4 讨论

（1）使用 CC-LIWC 分析文言文文本

将 CC-LIWC 词典用于分析孔子的《论语》（节选包括《学而》《为政》《八佾》《述而》）以及韩非的自传文章《孤愤》，得到两者词频差异最大词类前十位的统计结果如图 2 所示。这里词频的统计方法表示为公式 1。

$$词类A的词频=\frac{词类A的词数}{文章包含的所有LIWC词类词次总和}$$ （公式 1）

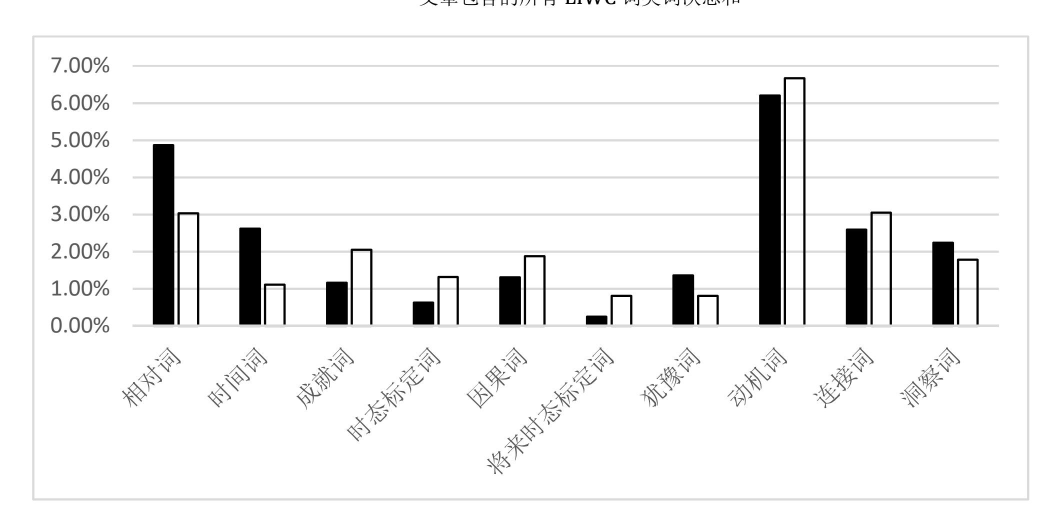


图 2 《论语（节选）》（左）与《孤愤》（右）LIWC 词频差异

从图 2 可以看出，《论语（节选）》有更多的“相对词”、“时间词”、“犹豫词”和“洞察词”，而韩非的《孤愤》则有更多的 “成就词”、“时态标定词”、“因果词”、“未来导向词”、“动机词”和“连接词”。其中，“时间词”是“相对词”的子类，“因果词”、“犹豫词”和“洞察词”是“认知过程词”的子类，“成就词”是“动机词”的子类，“将来时态标定词”是“时态标定词”的子类。由此可见，《论语（节选）》更关注时间的当下（某时某刻，某段时间等），认知上更偏好洞察（体会，了解，恍然大悟等），并对事物的认知所得结论始终保持着一定的余地（大约，好像），体现了儒家的中庸之道，而法家在时间上更关注未来，且重视个人成就动机，在认知上注重因果关系的分析，体现了法家重视辩证的思想。

（2）局限

由于文言文有其固定的写作格式，导致文言虚词（之乎者也等）使用非常频繁，且多为相同字词多次重复出现，造成统计结果中功能词的词频占比很大，但对于文言虚词目前在其心理学意义上还没有较为合适的解读方式。

其次是文言词中，一词多义、一词多性的情况普遍存在，对词典中词汇的分类存在一定影响。这也会影响最终分析结果的精确性。

5 总结

本研究主要介绍了古文 LIWC 词典 CC-LIWC 的构建过程，包括数据的获取、词汇的匹配和人工标注等，并介绍了最终生成的词典的词类构成占比情况。最后使用 CC-LIWC 词典对文

言文进行初步分析，结果表明，CC-LIWC 词典能够有效区分文言文的表达倾向。接下来的研究中，还要对词典进行效度验证，并将词典应用于更多历史研究主题，比如对古人的大五人格特征分析等。

参考文献:

[1]Pennebaker, J.W., Boyd, R.L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. Austin, TX: University of Texas at Austin. DOI: 10.15781/T29G6Z

[2]Pennebaker, J.W., Booth, R.J., Boyd, R.L., & Francis, M.E. (2015). Linguistic Inquiry and Word Count: LIWC2015. Austin, TX: Pennebaker Conglomerates (www.LIWC.net).

[3]张美杰,在线汉语词典[DB/OL].http://xh.5156edu.com/,2017-08-28

[4]梅常发,文言文字典[DB/OL].http://wyw.hwxnet.com/, 2010-06-04.

[5]mahavivo,汉语古典文本数据库[DB/OL].https://github.com/mahavivo/scripta-sinica,2018-02-03

[6]丁 帆, 赵普光. 历史的轨迹:中国现当代文学研究七十年的实证分析——以论题词词频的统计为中心 [J]. 文艺研究, 2019,9):55-68

[7]张信勇. LIWC:一种基于语词计量的文本分析工具[J]. 西南民族大学学报(人文社会科学版), 2015(04):108-111.

[8]吴嵩, 金盛华, 蔡頠, et al. 基于语言内容的谎言识别[J]. 心理科学进展, 2012, 20(3).

[9]Gottschalk, L.A. (1997). The unobtrusive measurement of psychological states and traits. In Text Analysis for the Social Sciences, (Carl W. Roberts, editor). 117-129.

[10]Tausczik, Y.R., & Pennebaker, J.W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology, 29, 24-54.

[11]Pennebaker, J.W. (2011). The Secret Life of Pronouns: What Our Words Say About Us. New York: Bloomsbury (www.secretlifeofpronouns.com)

[12]Mehl, M. R. (2006). Quantitative text analysis. Handbook of Multimethod Measurement in Psychology, 141-156.

[13]Krippendorff, K., & Bock, M. A. (2009). The Content Analysis Reader. Sage.

[14]Hirsh, J.B., & Peterson, J.B. (2009). Personality and language use in self-narratives. Journal of Personality in Research, 43, 524-527.

[15]Zhao N , Jiao D , Bai S , et al. Evaluating the Validity of Simplified Chinese Version of LIWC in Detecting Psychological Expressions in Short Texts on Social Network Services[J]. PLOS ONE, 2016, 11(6):1-15.